Calibration tests for count data

Wei, Wei ; Held, Leonhard

Abstract: Calibration, the statistical consistency of forecast distributions and observations, is a central requirement for probabilistic predictions. Calibration of continuous forecasts has been widely discussed, and significance tests are commonly used to detect whether a prediction model is miscalibrated. However, calibration tests for discrete forecasts are rare, especially for distributions with unlimited support. In this paper, we propose two types of calibration tests for count data: tests based on conditional exceedance probabilities and tests based on proper scoring rules. For the latter, three scoring rules are considered: the ranked probability score, the logarithmic score and the Dawid-Sebastiani score. Simulation studies show that all the different tests have good control of the type I error rate and sufficient power under miscalibration. As an illustration, we apply the methodology to weekly data on meningoccocal disease incidence in Germany, 2001–2006. The results show that the test approach is powerful in detecting miscalibrated forecasts.

DOI: https://doi.org/10.1007/s11749-014-0380-8

Noname manuscript No.
(will be inserted by the editor)

Calibration tests for count data

Wei Wei · Leonhard Held

Received: date / Accepted: date

Abstract Calibration, the statistical consistency of forecast distributions and observations, is a central requirement for probabilistic predictions. Calibration of continuous forecasts has been widely discussed, and significance tests are commonly used to detect whether a prediction model is miscalibrated. However, calibration tests for discrete forecasts are rare, especially for distributions with unlimited support. In this paper, we propose two types of calibration tests for count data: tests based on conditional exceedance probabilities and tests based on proper scoring rules. For the latter, three scoring rules are considered: the ranked probability score, the logarithmic score and the Dawid-Sebastiani score. Simulation studies show that all the different tests have good control of the type I error rate and sufficient power under miscalibration. As an illustration, we apply the methodology to weekly data on meningoccocal disease incidence in Germany, 2001-2006. The results show that the test approach is powerful in detecting miscalibrated forecasts.

Keywords Calibration test · Count data · Predictive distribution · Proper scoring rules

1 Introduction

In the statistical analysis of predictions, forecasts usually take the form of probability distributions. How to evaluate the performance of predictive distributions is an essential component in forecast research. There is a strand of

Wei Wei
Division of Biostatistics, Institute of Social and Preventive Medicine,
University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland
E-mail: wei.wei@uzh.ch

Leonhard Held
Division of Biostatistics, Institute of Social and Preventive Medicine,
University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland
E-mail: leonhard.held@ifspm.uzh.ch

work in the econometrics literature relevant to forecast evaluation (Diebold and Mariano, 1995; Harvey et al, 1998; Christoffersen, 1998; Diebold et al, 1998; Corradi and Swanson, 2006). Murphy and Winkler (1987) proposed a general framework for the evaluation of point forecasts and called for the consideration of the joint distribution of forecast and observation. Gneiting et al (2007) contended that the goal of probabilistic forecasting is to *maximize the sharpness of the predictive distributions subject to calibration*. In this context, calibration refers to the statistical consistency between the probabilistic forecasts and the actual observations. Sharpness refers to the concentration of the predictive distributions.

Discussion of calibration falls into two general classes. The first class refers to calibration of continuous forecasts. The forecaster must report a probability density function across the possible values of such uncertain quantities. The second class concerns the calibration of probabilities of discrete forecasts. These include probabilities for a binary outcome "Yes/No" (e.g. whether it will rain tomorrow) and integer-valued outcome (e.g. how many people get infected).

For continuous forecasts, the probability integral transform (PIT) histogram (Dawid 1984, Gneiting et al 2007) is commonly used to assess calibration. Several tests based on proper scoring rules are proposed in Held et al (2010). Alternatively, Mason et al (2007) suggest the usage of the conditional exceedance probability (CEP) in a logistic regression framework to assess calibration of continuous probabilistic forecasts.

There are many methods to assess calibration of categorical forecasts. The Brier (Brier, 1950; Spiegelhalter, 1986) and the logarithmic score (Cox, 1958) can be used to assess calibration of binary forecasts. The Brier score has been extended to ordered multicategorical outcomes with a finite number of support points (Epstein, 1969). Czado et al (2009) modified the probability integral transform (PIT) histogram and discussed proper scoring rules for count forecasts. However, statistical hypothesis tests to assess calibration of count forecasts are rarely discussed.

In this paper, we fill this gap and provide several tools to assess calibration of statistical predictions of count data. Count data are commonly met in quantitative sciences, for example in econometrics, climate, ecology, finance, epidemiology and other areas (McCabe and Martin, 2005; Elsner and Jagger, 2006; Nelson and Leroux, 2006; Winkelmann, 2008; Frühwirth-Schnatter et al, 2009; Steyerberg, 2009; McCabe et al, 2011; Hilbe, 2011). Our specific motivation for this work comes from the analysis of surveillance data on infectious diseases. Here disease cases are notified in surveillance registries and reported as (daily or weekly) counts of disease incidence. One of the main tasks of such registries is to flag a warning if disease incidence is rising. The conventional approach to do this is to compute a probabilistic one-step-ahead prediction based

on a simple regression model applied to historical data (Farrington et al 1996; Heisterkamp et al 2006; Noufaily et al 2013; Manitz and Höhle 2013). If the observed counts exceed a pre-specified threshold, for example the 99%-quantile of the predictive distribution, then an alarm is flagged. Validation of such an *outbreak detection* procedure is typically based on extensive simulation studies where certain operation characteristics, such as the false positive rate and the probability that an outbreak is detected, are evaluated. However, an inherent problem of such an approach is that it implicitly assumes that the historical records do not contain outbreaks, otherwise those have to be down-weighted using iterative procedures. In contrast, a model-based *outbreak prediction* approach allows for past outbreaks and provides a potential alternative (Held et al 2006). Here the idea is to fit a fairly realistic model to the time series at hand, e.g. based on recent developments in modelling infectious disease counts (Held et al 2005; Paul et al 2008; Held and Paul 2012). A warning for increasing disease incidence will then be flagged if a pre-specified upper quantile of the one-step-ahead prediction interval exceeds a certain limit. It is central for such a *model-based* approach that the predictions are well calibrated. This can be investigated by applying the methodology developed in this paper to the one-step-ahead forecasts for the data at hand. An example is given in Section 4.2.

The outline of the paper is as follows: In Section 2, we derive specific forms of the proper scoring rules considered for Poisson and negative binomial predictions: the ranked probability score (RPS) (Epstein 1969, Gneiting et al 2007), the logarithmic score (LS) (Good, 1952) and the Dawid-Sebastiani score (DSS) (Dawid and Sebastiani, 1999). In Section 3, we first introduce the CEP regression test, and adapt it to count forecasts. We also develope two types of calibration tests based on proper scoring rules: unconditional and regression tests. For the latter, certain approximations are required and we prove that the approximation error can be bounded at any pre-specified level. Results based on simulated and real data are presented in Section 4.

2 Proper scoring rules

Scoring rules assign numerical scores to probabilistic forecasts and can be viewed as penalties on the difference between observations and predictions. A scoring rule is proper if the expected value of the score is minimised if the prediction is ideal, that is, the observation is from the predictive distribution. Following the terminology of Gneiting et al (2007), such a prediction is called ideal, perfect or strongly calibrated. It is strictly proper if the minimum is unique(Gneiting et al 2007). Strict propriety ensures that both calibration and sharpness are being addressed (Winkler, 1996; Czado et al, 2009).

Three different types of scoring rules are considered. The logarithmic score (LS) is the negative log-likelihood evaluated at the actual observation y_{obs},

i.e. $\text{LS}(y_{obs}) = -\log f(y_{obs})$. The Dawid-Sebastiani score $\text{DSS}(y_{obs}) = \tilde{y}_{obs}^2 + \log \sigma^2$, where $\tilde{y}_{obs} = (y_{obs} - \mu)/\sigma$, depends only on the mean μ and the variance σ^2 of the predictive distribution. Finally, the ranked probability score $\text{RPS}(y_{obs}) = \sum_{t=0}^{\infty} \{F(t) - \mathbf{1}(y_{obs} \leq t)\}^2$ is the sum of the Brier scores (Brier 1950) for binary predictions at all possible thresholds t. This has been suggested for data with more than two ordinal categories (Epstein, 1969). It can be written as

$$\text{RPS}(y_{obs}) = \mathsf{E}\,|Y - y_{obs}| - \mathsf{E}\,|Y - Y'|/2 \tag{1}$$

where Y and Y' are independent and identically distributed according to the predictive distribution (Gneiting et al 2007).

Let $f(.)$ and $F(.)$ denote the probability mass function and cumulative distribution function, specifically, $f(.;\mu)$; $F(.;\mu)$ for the Poisson distribution $\text{Po}(\mu)$, and $f(.;\mu,\psi)$; $F(.;\mu,\psi)$ for the negative binomial distribution $\text{NBin}(\mu,\psi)$. Here $\mu > 0$ denotes the mean of each distribution whereas $\psi > 0$ accounts for overdispersion of the negative binomial distribution. More specifically, the probability mass function of the negative binomial distribution $\text{NBin}(\mu,\psi)$ is

$$f(y;\mu,\psi) = \frac{\Gamma(y+\psi)}{y!\Gamma(\psi)} \left(\frac{\mu}{\mu+\psi}\right)^y \left(\frac{\psi}{\psi+\mu}\right)^\psi \quad \text{for } y = 0,1,2,\ldots,$$

here $\Gamma(\cdot)$ denotes the gamma function. For $\psi \to \infty$, the variance $\sigma^2 = \mu + \mu^2/\psi$ converges to the mean μ and the negative binomial will become a Poisson distribution.

Table 1 lists analytic formulas for the proper scoring rules considered based on Poisson and negative binomial predictions. The RPS formula for a Poisson prediction and a negative binomial prediction is derived in the Appendix A. The RPS formula for a negative binomial prediction involves an infinite sum. This term turns out to be the expectation of $|Y - Y'|/2$, the second term in Equation (1). In practice this infinite sum is computed by truncation at a sufficiently accurate value, see Section 3.4 for details.

3 Calibration tests

3.1 CEP regression test

Let $q(p)$ denote the p-quantile of the predictive distribution P. The conditional exceedance probability (CEP) for the p-quantile $q(p)$ is defined as

$$\mathsf{Pr}\{Y_{obs} > q(p)\}. \tag{2}$$

For an ideal continuous forecast, *i.e.* Y_{obs} is distributed according to P, it equals $1 - p$ for any fixed proportion $p \in (0,1)$ and is independent of $q(p)$.

Table 1: Formulas of proper scoring rules for Poisson Po(μ) and negative binomial NBin(μ, ψ) forecasts.

Forecast	Proper scoring rule
Po(μ)	$\text{LS}(y_{obs}) = \mu - y_{obs} \log \mu + \log(y_{obs}!)$
Po(μ)	$\text{DSS}(y_{obs}) = \frac{(y_{obs}-\mu)^2}{\mu} + \log \mu$
Po(μ)	$\text{RPS}(y_{obs}) = (y_{obs} - \mu)\{2F(y_{obs}; \mu) - 1\}$ $+2\mu f(y_{obs}; \mu) - \mu e^{-2\mu}\{I_0(2\mu) + I_1(2\mu)\}$ where $I_m(x)$ is the Bessel function of the first kind.
NBin(μ, ψ)	$\text{LS}(y_{obs}) = \log(y_{obs} + \psi) + \log B(y_{obs} + 1, \psi)$ $+y_{obs} \log \frac{\mu+\psi}{\mu} + \psi \log \frac{\mu+\psi}{\psi}$, where $B(x, y)$ is the Beta function.
NBin(μ, ψ)	$\text{DSS}(y_{obs}) = \frac{(y_{obs}-\mu)^2}{\mu(1+\mu/\psi)} + \log\{\mu(1 + \mu/\psi)\}$
NBin(μ, ψ)	$\text{RPS}(y_{obs}) = y_{obs}\{2F(y_{obs}; \mu, \psi) - 1\}$ $+\mu\{1 - 2F(y_{obs} - 1; \mu(1 + 1/\psi), \psi + 1)\}$ $+ \sum_{k=0}^{\infty} \sum_{j=0}^{k-1} (k - j) f(k; \mu, \psi) f(j; \mu, \psi)$.

Mason et al (2007) propose a logistic regression approach to test for miscalibration. More specifically, suppose a model gives a set of independent predictive distributions P_i $(i = 1, 2, \ldots, n)$, and the corresponding observations are denoted as $y_{obs,i}$. Let the binary indicator $w_i(p) = I_{\{y_{obs,i} > q_i(p)\}}$ be the response variable in a logistic regression with explanatory variable $q_i(p)$:

$$\text{logit}[\mathsf{Pr}\{Y_{obs,i} > q_i(p)\}] = \beta_0 + \beta_1 q_i(p). \tag{3}$$

Fitting a logistic regression model gives estimates of β_0 and β_1. Under the null hypothesis that P_i's are well calibrated, we have $\beta_0 = \text{logit}(1 - p)$ and $\beta_1 = 0$. Mason et al (2007) suggest to use a test of the null hypothesis H_0: $\beta_1 = 0$. Alternatively, Held et al (2010) propose to consider H_0: $\beta_0 = \text{logit}(1 - p), \beta_1 = 0$ with a likelihood ratio or Wald test.

However, for count data, $\mathsf{Pr}\{Y_{\text{obs}} > q(p)\}$ does not equal $1 - p$ any more. The exceedance probability (2) now depends on the predictive distribution. For example, consider two forecasts: a Poisson forecast Po(5) and a normal forecast $\text{N}(\mu = 5, \sigma^2)$ with $\sigma^2 > 0$. The median is $q(0.5) = 5$ for both forecasts, but the corresponding exceedance probability $\mathsf{Pr}\{Y_{obs} > q(p)\}$ is 0.384 for the Poisson and 0.5 for the normal prediction. Therefore, the regression test based on (3) is no longer valid.

To remedy this, let us denote $\mathsf{Pr}\{Y_{obs} > q(p)\} = 1 - p^*$. Then we have

$$\begin{aligned}\text{logit}[\mathsf{Pr}\{Y_{obs,i} > q_i(p)\}] &= \text{logit}(1 - p_i^*) \\ &\doteq o_i.\end{aligned}$$

Note that the predictive distribution is assumed to be entirely known, therefore p_i^* can be computed. Using an offset $o_i = \text{logit}(1 - p_i^*)$, the logistic regression model (3) can be adjusted accordingly:

$$\text{logit}[\mathsf{Pr}\{Y_{obs,i} > q_i(p)\}] = \beta_0 + \beta_1 q_i(p) + o_i. \tag{4}$$

Therefore, we still can test H_0: $\beta_1 = 0$ in the logistic regression (4) (Mason et al, 2007). Alternatively, a likelihood ratio test for H_0: $\beta_0 = 0, \beta_1 = 0$ (Held et al, 2010) can also be conducted based on the regression estimates from model (4).

3.2 Score tests

3.2.1 Unconditional tests

Suppose a sequence of score values s_i $(i = 1, 2, \ldots, n)$ has been computed based on each observation y_i and prediction P_i. Score value s_i here can be either RPS, LS or DSS. Using the mean score $\bar{s} = \sum_{i=1}^{n} s_i/n$, an asymptotically standard normal distributed test statistic can be conducted without any distribution assumption on the scores s_i (Spiegelhalter 1986, Held et al 2010). The central limit theorem of Liapounov applies to a sequence of independent random variables that are not necessarily identically distributed. Therefore, no distribution assumption of the scores s_i is required here. The theorem requires that the third moment of each s_i exists (DeGroot and Schervish, 2012), which can be proved for all three scores we considered. The test statistic takes the form

$$Z_s = \frac{\bar{s} - \mathsf{E}_0(\bar{s})}{\mathsf{Var}_0(\bar{s})^{1/2}}, \tag{5}$$

where $\mathsf{E}_0(\bar{s})$ and $\mathsf{Var}_0(\bar{s})$ are expectation and variance of the mean scores $\bar{s}$ under the null hypothesis. Usually a two-sided p-values is computed based on the value of Z_s.

3.2.2 Score regression

Held et al (2010) propose a regression approach based on the scores s_i using the expectation $\mathsf{E}_0(s_i)$ under the null hypothesis,

$$s_i = c + d \cdot \mathsf{E}_0(s_i) + \epsilon_i, \tag{6}$$

where the errors ϵ_i have mean zero, but are not necessarily normal. For an ideal forecast we have $c = c_0 = 0$ and $d = d_0 = 1$, so we can test the null hypothesis H_0: $c = c_0$, $d = d_0$ using this regression. A heteroscedastic model

should be used, if $\mathsf{Var}_0(s_i)$ is not constant. This is accomplished by using the weights $1/\mathsf{Var}_0(s_i)$ in the regression model (6).

To assess the null hypotheses H_0: $c = c_0$, $d = d_0$, one can perform a standard significance test. Let $\widehat{V}$ denote the estimated variance-covariance matrix of the (weighted) least squares estimates $(\hat{c}, \hat{d})^T$ based on model (6), we can calculate

$$T_s = (\hat{c} - c_0, \hat{d} - d_0)V^{-1}(\hat{c} - c_0, \hat{d} - d_0)^T, \tag{7}$$

which, for an ideal forecast, is asymptotically χ^2-distributed with 2 degrees of freedom. The score regression tests used in Section 4 are based on this approach.

We can also test the two coefficients separately. For example, one can consider the reduced null hypothesis H_0: $c = c_0 = 0$ and use the squared t-statistic $\widetilde{T}_s = (\hat{c} - c_0)^2/\operatorname{se}(\hat{c})^2$, here $\operatorname{se}(\hat{c})$ denotes the standard error of $\hat{c}$. Under the null hypothesis of an ideal forecast, $\widetilde{T}_s$ is asymptotically χ^2-distributed with one degree of freedom.

3.3 Computation of E_0 and Var_0

For an ideal forecast, we assume that the data-generating distribution of Y_{obs} equals the forecast distribution P. Expectation E_0 and variance Var_0 of the scores for ideal forecasts need to be computed for application of the tests. Generally, it is difficult to get analytic formulas of E_0 and Var_0, both for Poisson and negative binomial predictions. Informally, the order of difficulty is:

$$\begin{aligned} \mathsf{E}_0(\cdot) &\preceq \mathsf{Var}_0(\cdot), \\ \mathrm{DSS} \preceq \mathrm{LS} &\preceq \mathrm{RPS}, \\ \mathrm{Po}(\mu) &\preceq \mathrm{NBin}(\mu, \psi). \end{aligned}$$

Both $\mathsf{E}_0(\mathrm{DSS})$ and $\mathsf{Var}_0(\mathrm{DSS})$ can be computed analytically both for Poisson and negative binomial predictions, as well as $\mathsf{E}_0(\mathrm{RPS})$ for Poisson predictions. For a $\mathrm{Po}(\mu)$ prediction, $\mathsf{E}_0(\mathrm{DSS}) = 1 + \log\mu$ and $\mathsf{Var}_0(\mathrm{DSS}) = 2 + 1/\mu$; whereas, $\mathsf{E}_0(\mathrm{DSS}) = 1 + \log(\mu + \mu^2/\psi)$ and $\mathsf{Var}_0(\mathrm{DSS}) = 2 + 6/\psi + 1/(\mu + \mu^2/\psi)$ for a $\mathrm{NBin}(\mu, \psi)$ prediction. When ψ goes to infinity, the variance $\mathsf{Var}_0(\mathrm{DSS})$ of a negative binomial prediction converges to the variance of a Poisson prediction. For a Poisson prediction with increasing mean μ, the variance converges to 2 as it should, since this is the variance of DSS for a normal prediction (Held et al, 2010).

$\mathsf{E}_0(\mathrm{RPS})$ are difficult to calculate due to the term $\mathsf{E}\,|Y - Y'|/2$ in Equation (1). Viewed as a function of y_{obs}, we have $\mathrm{RPS}(y_{obs}) = \mathsf{E}(Z \,|\, Y_{obs} = y_{obs}) - \mathsf{E}\,|Y - Y'|/2$, where $Z = |Y - Y_{obs}|$. The expectation of the first component is $\mathsf{E}\,\mathsf{E}(Z \,|\, Y_{obs} = y_{obs}) = \mathsf{E}(Z) = \mathsf{E}\,|Y - Y_{obs}|$. Therefore we have

$\mathsf{E}_0(\text{RPS}) = \mathsf{E}\,|Y - Y'|/2$. Using a result from Katti (1960), we obtain for a Poisson distributed prediction

$$\mathsf{E}_0(\text{RPS}) = \mu e^{-2\mu}\{I_0(2\mu) + I_1(2\mu)\}, \tag{8}$$

where $I_m(x)$ is the Bessel function of the first kind, compare Table 1. However, it is difficult to get an explicit formula for the variance of a Poisson prediction. For a negative binomial prediction, both expectation and variance of RPS are in general not available analytically. An approximate approach to compute E_0 and Var_0 of LS and RPS is discussed in the following section.

3.4 Numerical computation of E_0 and Var_0

In what follows we describe mathematical results useful for the numerical computation of E_0 and Var_0 of LS and RPS. These results ensure that the approximation error, defined as the absolute difference between the approximate and the true value, can not exceed a pre-specified limit δ. For example, if $\mathsf{E}_0^*(\text{LS})$ denotes the approximation of $\mathsf{E}(\text{LS})$, the approximation error $|\mathsf{E}(\text{LS}) - \mathsf{E}_0^*(\text{LS})|$ should be smaller than the pre-specified limit δ. In the applications described in Section 4, we set the upper limit to $\delta = 10^{-4}$.

3.4.1 Logarithmic score

Based on the formulas given in Table 1, expectation and variance of LS can be calculated as follows: For a $\text{Po}(\mu)$ prediction, we obtain

$$\mathsf{E}_0(\text{LS}) = \mu - \mu\log\mu + e^{-\mu}\sum_{k=1}^{\infty}\frac{\mu^k \log(k!)}{k!} \text{ and} \tag{9}$$

$$\mathsf{Var}_0(\text{LS}) = \sum_{k=1}^{\infty}(-k\log\mu + \log k!)^2 f(k;\mu) - \{\mathsf{E}_0(\text{LS}) - \mu\}^2. \tag{10}$$

For a $\text{NBin}(\mu,\psi)$ prediction, we obtain

$$\begin{aligned}\mathsf{E}_0(\text{LS}) = &\sum_{k=0}^{\infty}\{-\log\Gamma(k+\psi) + \log\Gamma(k+1)\}f(k;\mu,\psi)\\ &- \psi\log\psi - \mu\log\mu + (\psi+\mu)\log(\psi+\mu) + \log\Gamma(\psi) \text{ and}\end{aligned} \tag{11}$$

$$\begin{aligned}\mathsf{Var}_0(\text{LS}) = &\sum_{k=0}^{\infty}\left\{-\log\Gamma(k+\psi) + \log\Gamma(k+1) + k\log\frac{\mu+\psi}{\mu}\right\}^2 f(k;\mu,\psi)\\ &- \{\mathsf{E}_0(\text{LS}) + \psi\log\psi - \psi\log(\psi+\mu) - \log\Gamma(\psi)\}^2.\end{aligned} \tag{12}$$

It is natural to approximate $\mathsf{E}_0(\text{LS})$ of a Poisson prediction by truncating the infinite sum in Equation (9) at some upper value K_1, we say. This defines the approximation $\mathsf{E}_0^*(\text{LS})$ of $\mathsf{E}_0(\text{LS})$. Similarly we approximate $\mathsf{Var}_0(\text{LS})$ by truncating the infinite sum in Equation (10) at some upper value K_2. This gives

the approximation $\mathsf{Var}_0^*(\mathrm{LS})$ of $\mathsf{Var}_0(\mathrm{LS})$. For negative binomial predictions, the infinite sums in equation (11) and (12) are truncated at the upper values K_3 and K_4 to obtain the approximations $\mathsf{E}_0^*(\mathrm{LS})$ and $\mathsf{Var}_0^*(\mathrm{LS})$, respectively.

Theorem 1 *Let $q(p;\mu)$ and $q(p;\mu,\psi)$ denote the p-quantile of the* $\mathrm{Po}(\mu)$ *and the* $\mathrm{NBin}(\mu,\psi)$ *distribution, respectively. Fix $\delta > 0$.*

(a) With $K_1 = q(1-\delta/(\mu^2+3\mu+1);\mu)+2$, the approximation $\mathsf{E}_0^*(\mathrm{LS})$ of Equation (9) has approximation error smaller than δ;
(b) With $K_2 = q(1-\delta/g(\mu);\mu)+3$, the approximation $\mathsf{Var}_0^*(\mathrm{LS})$ of Equation (10) has approximation error smaller than δ, where $g(\mu) = \mu^3+6\mu^2+7\mu+1$;
(c) With $K_3 = g_1(\delta,\mu,\sigma)$ as given in the Appendix, the approximation $\mathsf{E}_0^*(\mathrm{LS})$ of Equation (11) has approximation error smaller than δ;
(d) With $K_4 = g_2(\delta,\mu,\sigma)$ as given in the Appendix, the approximation $\mathsf{Var}_0^*(\mathrm{LS})$ of Equation (12) has approximation error smaller than δ.

Proof The proof will be given in the Appendix B.

This theorem implies that for any pre-specified limit δ, we can find the corresponding values K_1, K_2, K_3 or K_4 to control the approximation error within δ.

For a $\mathrm{Po}(\mu)$ prediction, $\mathsf{E}_0(\mathrm{LS})$ can also be viewed as Shannon entropy, defined as

$$-\sum_{k=1}^{\infty} f(k;\mu)\log\{f(k;\mu)\}.$$

Knessl (1998) proposes a simple representation of this entropy:

$$\mathsf{E}_0(\mathrm{LS}) = \frac{1}{2}+\frac{1}{2}\log(2\pi\mu)-\frac{1}{12\mu}-\frac{1}{24\mu^2}-\frac{19}{360\mu^3}+O(1/\mu^4). \qquad (13)$$

This representation works well for large μ with an approximation error of order $1/\mu^4$. Therefore, a modified approximation of $\mathsf{E}_0(\mathrm{LS})$ based on Theorem 1 and Equation (13) can be used, which avoids calculation of the truncated sum if μ is large, say $\mu > \mu_0$:

$$\mathsf{E}_0^*(\mathrm{LS}) = \begin{cases} \mu-\mu\log\mu+e^{-\mu}\sum_{k=1}^{K_1}\frac{\mu^k\log(k!)}{k!} & \text{if } \mu \le \mu_0\ , \\ \frac{1}{2}+\frac{1}{2}\log(2\pi\mu)-\frac{1}{12\mu}-\frac{1}{24\mu^2}-\frac{19}{360\mu^3} & \text{if } \mu > \mu_0. \end{cases}$$

In Section 4 we use $\mu_0 = 10$ which ensures that $|\,\mathsf{E}_0^*-\mathsf{E}_0\,|$ is always of the order 10^{-4}.

3.4.2 Ranked probability score

For a Po(μ) prediction, the expectation $\mathsf{E}_0(\text{RPS})$ can be calculated analytically by Formula (8) in Section 3.3. For the variance under the null hypothesis, we will use the approximation:

$$\begin{aligned}\mathsf{Var}_0^*(\text{RPS}) = &\sum_{k=0}^{K_1^*}[(k-\mu)\{2F(k;\mu)-1\}+2\mu f(k;\mu)]^2 f(k;\mu) \\ &- 4\mu^2 e^{-4\mu}\{I_0(2\mu)+I_1(2\mu)\}^2. \end{aligned} \tag{14}$$

For a NBin(μ, ψ) prediction, the expectation $\mathsf{E}_0(\text{RPS})$ can be computed with the hypergeometric function ${}_2F_1$ (Katti, 1960), if $4\mu(1+\mu/\psi)/\psi < 1$:

$$\mathsf{E}_0(\text{RPS}) = \mu(1+\mu/\psi)\ {}_2F_1(1+\psi, 1/2; 2; -4\mu(1+\mu/\psi)/\psi).$$

However, this formula is not valid when $4\mu(1+\mu/\psi)/\psi \geq 1$, due to non-convergence of the hypergeometric function. In this case, we will approximate $\mathsf{E}_0(\text{RPS})$ by

$$\mathsf{E}_0^*(\text{RPS}) = \sum_{k=0}^{K_2^*}\sum_{j=0}^{k-1}(k-j)f(k;\mu,\psi)f(j;\mu,\psi). \tag{15}$$

For $\mathsf{Var}_0(\text{RPS})$, we will use

$$\begin{aligned}\mathsf{Var}_0^*(\text{RPS}) = &\sum_{k=0}^{K_3^*}[\mu\{1-2F(k-1;\mu(1/\psi+1),\psi+1)\} \\ &+ k\{2F(k;\mu,\psi)-1\}]^2 f(k;\mu,\psi) - 4\,\mathsf{E}_0^*(\text{RPS})^2. \end{aligned} \tag{16}$$

Theorem 2 *Fix $\delta > 0$.*

(a) For $K_1^* = \max[q\{1-\delta/(10\mu^2+\mu);\mu\}+2, \exp(2)]$, the error of the approximation (14) is smaller than δ;

(b) For $K_2^* = \max[q\{1-\delta/\mu;\mu(1+1/\psi),\psi+1\}+1, \exp(2)]$, the error of the approximation (15) is smaller than δ;

(c) For

$$\begin{aligned}K_3^* = \max[&q\{\delta/l_5;\mu(1+2/\psi),\psi+2\}+2, q(\delta/l_5;\mu,\psi), \exp(2), \\ &q\{\delta/l_5;\mu(1+1/\psi),\psi+1\}+1, K_2^*]\end{aligned}$$

where $l_5 = \mu^2+2\mu+2$, the error of the approximation (16) is smaller than δ.

Proof The proof will be given in the Appendix B.

4 Applications

4.1 Simulation data

A good test should be able to control the type I error and have sufficient power to detect deviations from the hypothesis. To assess the type I error and the power of the different tests, we simulate 10,000 datasets with different number $n \in \{10, 20, 50, 100, 500\}$ of Poisson or negative binomial predictions. This is done both for ideal and miscalibrated forecasts. The type I error and the power are assessed by the proportion of rejected null hypotheses. The Monte Carlo standard error of these proportions is always smaller than 0.01. The significance level is always set to 5%.

4.1.1 Type I error control

Suppose that the true data-generating distribution is $Y_i \sim \text{Po}(\mu_i)$ or $\text{NBin}(\mu_i, \psi_i)$ (setting $\psi_i = \mu_i$), where the mean μ_i is a realization of a gamma $\text{G}(10, 0.5)$ random variable with mean 20 and variance 40. The ideal forecast P_i is equal to the data-generating distribution Y_i, i.e. $P_i = Y_i$.

Table 2 provides the proportion of rejected null hypotheses for ideal forecasts. A test with good Type I error control should have an proportion of rejected null hypotheses of around 5%. In this study, all unconditional score and CEP regression tests perform quite well even for a small number n of observations. The score regression test produces higher rejection rates when the number of observations is smaller than 50, but the rates decrease to 0.05 for larger sample sizes ($n = 100$).

4.1.2 Power assessment: miscalibrated forecasts with a different location or scale

In the statistical forecast analysis, the data generating distribution is hardly known. A series of forecasts from one model may be with different locations (means) or scales (variances) comparing to the data-generators. To assess the power of these tests in detecting miscalibrated forecasts, we set different values to either the location or the scale of the observation generator Y_i and prediction P_i, respectively. For simplicity we always use $\psi_i = \mu_i$, where μ_i is independently sampled from a gamma distribution $\text{G}(10, 0.5)$.

To detect miscalibration with different locations, the observations and forecast distributions are defined as follows:

$$\begin{cases} Y_i \sim \text{Po}(\mu_i) \\ P_i \sim \text{Po}(\mu_i \pm 0.3\mu_i) \end{cases} \quad \text{or} \quad \begin{cases} Y_i \sim \text{NBin}(\mu_i, \psi_i) \\ P_i \sim \text{NBin}(\mu_i \pm 0.3\mu_i, \psi_i). \end{cases}$$

Table 2: Proportion of null hypothesis rejection for ideal forecasts.

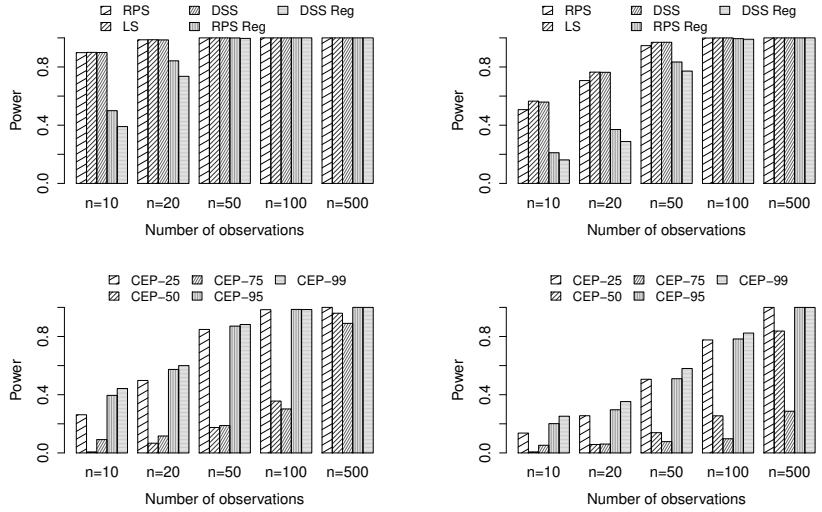
Test	**Poisson**					**Negative binomial**				
	n=10	n=20	n=50	n=100	n=500	n=10	n=20	n=50	n=100	n=500
Unconditional										
RPS	0.043	0.046	0.046	0.053	0.050	0.049	0.051	0.051	0.057	0.052
LS	0.045	0.046	0.046	0.051	0.051	0.044	0.044	0.046	0.051	0.047
DSS	0.044	0.045	0.046	0.052	0.051	0.047	0.044	0.045	0.048	0.049
Score regression										
RPS	0.165	0.110	0.073	0.064	0.056	0.172	0.114	0.083	0.074	0.056
DSS	0.192	0.132	0.088	0.073	0.058	0.200	0.140	0.096	0.082	0.055
CEP										
CEP-25%	0.024	0.024	0.041	0.046	0.051	0.022	0.028	0.041	0.044	0.054
CEP-50%	0.013	0.035	0.042	0.051	0.048	0.012	0.040	0.048	0.047	0.050
CEP-75%	0.024	0.030	0.039	0.046	0.051	0.029	0.029	0.040	0.045	0.049
CEP-95%	0.043	0.037	0.037	0.031	0.046	0.048	0.042	0.033	0.035	0.044
CEP-99%	0.059	0.054	0.047	0.039	0.036	0.069	0.048	0.047	0.039	0.033

To detect miscalibrated forecasts with different scales or variances, the performance of each test is explored in the following setting:

$$\begin{cases} Y_i \sim \text{Po}(\mu_i) \\ P_i \sim \text{NBin}(\mu_i, \psi_i) \end{cases} \quad \text{or} \quad \begin{cases} Y_i \sim \text{NBin}(\mu_i, \psi_i) \\ P_i \sim \text{Po}(\mu_i). \end{cases}$$

We choose negative binomial distribution as the counterpart of Poisson distribution, and vice versa. In this setting, the mean of the forecast distribution is the same as the mean of the observation generator, while the variance is twice or half as large as the variance of the generator to reflect under or overdispersion, respectively.

Figure 1 and Figure 2 display the proportion of rejected null hypotheses for each test in the different scenarios. This proportion can be interpreted as the power to detect a miscalibrated prediction: the larger the value, the better the test performs. Overall, the power of the tests increases with increasing number of observations n. Tests based on proper scoring rules work better than the CEP approach, and reach 100% power already for $n = 100$ observations. Linear model asymptotics ensure that the estimates are consistent even if the error terms ϵ_i are not necessarily normal (Held et al, 2010). This explains why regression tests are powerful although the distribution of proper scoring rules is unknown. Among them, the unconditional score tests work better than the score regression tests for very low number of observations. For detecting miscalibration with different scales, see Figure 2, unconditional tests work better for underdispersed forecasts (Figure 2a), but not for overdispersed forecasts (Figure 2b). Note that for miscalibrated forecasts with correct locations but different scales, s_i, $\mathsf{E}_0(s_i)$ and $\mathsf{Var}_0(s_i)$ are all larger for a more dispersed fore-



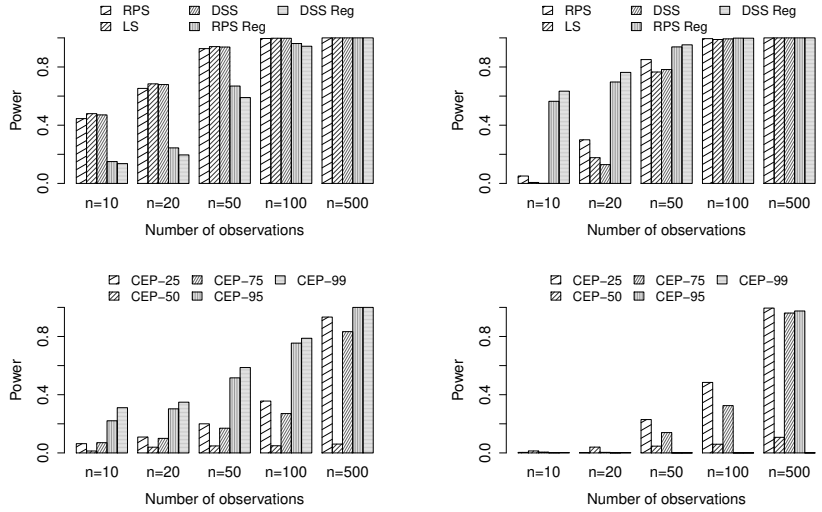
(a) Miscalibrated Poisson prediction with different location

(b) Miscalibrated negative binomial prediction with different location

Fig. 1: Power of calibration tests for miscalibration with different location.

cast. Therefore, the standard error of $\overline{s}$ is larger for negative binomial forecasts than Poisson forecasts and it is more difficult to reject the null hypothesis for negative binomial forecasts. In contrast, the elements of the covariance matrix of the (weighted) least squares estimate V in Equation (7) are not always larger for negative binomial forecasts, so the power of the unconditional tests is lower in this case (Figure 2b).

The CEP tests perform slightly worse than the tests based on proper scoring rules. Among them, tests based on large quantiles (95% and 99%) work better than those based on 50% or 75%- quantiles. However, note that for large p, the logistic regression of CEP might fail due to only zero responses $w_i(p)$ in Equation (4). In this case the maximum likelihood (ML) estimates diverge, and the result of the CEP test has been set to "no rejection". This explains why the CEP test based on 99%-quantile does not reach any reasonable power even for large n, see Figure 2b. The CEP test based on 50%-quantiles shows extreme low power when detecting miscalibration with different variance. This indicates that the CEP approach based on 50%-quantiles is not sensitive to miscalibrated forecasts with different variances. All tests work better for Poisson than for negative binomial predictions, which is due to the larger variance of the negative binomial distribution.



(a) Observations are from negative binomial distributions with miscalibrated Poisson predictions

(b) Observations are from Poisson distributions with miscalibrated negative binomial predictions

Fig. 2: Power of calibration tests for miscalibration with different scale.

4.2 Meningococcal disease incidence in Germany

One important characteristic in statistical analysis of counts is overdispersion. Therefore a negative binomial model is often proposed to accommodate it. Paul et al (2008) compared Poisson and negative binomial models with or without an autoregressive component for weekly reported cases of meningococcal disease incidence in Germany, 2001–2006. The formulation has been extended in Held and Paul (2012) to allow for seasonal variation in the autoregressive component.

More specifically, suppose that the number of disease cases Y_t in week t is assumed either Poisson $\mathrm{Po}(\mu_t)$ or negative binomial $\mathrm{NBin}(\mu_t, \psi)$ distributed with mean

$$\begin{aligned} \mu_t &= \nu_t && \text{(model A)}, \\ \mu_t &= \lambda y_{t-1} + \nu_t && \text{(model B)}, \\ \text{or } \mu_t &= \lambda_t y_{t-1} + \nu_t && \text{(model C)}. \end{aligned}$$

Here y_{t-1} are the observed number of cases in the previous week $t-1$ and ν_t includes sinusoidal terms to account for seasonality in disease incidence:

$$\log \nu_t = \alpha_\nu + \gamma_\nu \sin(2\pi t/52) + \delta_\nu \cos(2\pi t/52).$$

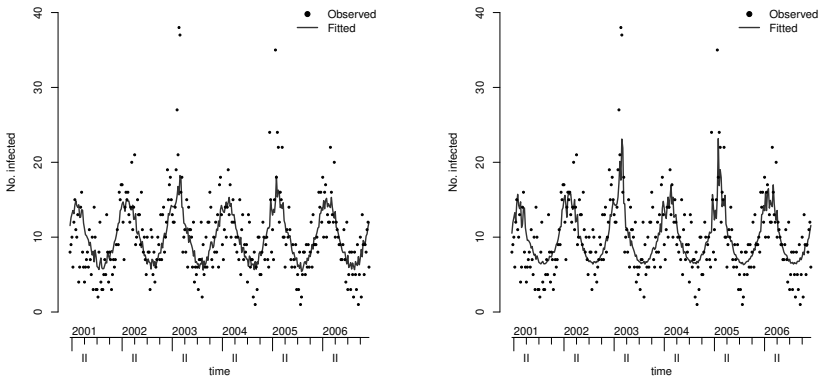
In model C, the autoregressive parameter λ_t is also allowed to show seasonal variation:

$$\log \lambda_t = \alpha_\lambda + \gamma_\lambda \sin(2\pi t/52) + \delta_\lambda \cos(2\pi t/52).$$

Note that model B and C account for autocorrelation by incorporating disease counts y_{t-1} of the previous week whereas all observations are assumed to be independent in model A. Figure 3 displays the data and the fit from the negative binomial models B and C. It is worth noting that model C captures the large counts at the beginning of 2003 and 2005 better than model B.

Table 3: Mean scores and p-values of Poisson (Poi) or negative binomial (NBin) models A and B for data on meningococcal disease incidence in Germany.

Test	**Poi A**		**Poi B**		**Poi C**		**NBin A**		**NBin B**		**NBin C**	
	Mean	p-value	Mean	p-value	Mean	p-value	Mean	p-value	Mean	p-value	Mean	p-value
Unconditional												
RPS	2.27	<0.001	2.29	<0.001	2.26	<0.001	2.26	0.83	2.25	0.58	2.23	0.45
LS	4.07	<0.001	3.95	<0.001	3.89	<0.001	3.85	0.86	3.79	0.50	3.76	0.41
DSS	2.86	<0.001	2.85	<0.001	2.82	<0.001	2.77	0.53	2.77	0.31	2.76	0.25
Score regression												
RPS		0.006		0.003		0.003		0.84		0.84		0.80
DSS		0.006		0.004		0.003		0.46		0.57		0.52
CEP												
CEP-10%		0.25		0.009		0.003		0.13		0.52		0.60
CEP-25%		0.31		0.16		0.084		0.17		0.97		0.55
CEP-50%		0.62		0.28		0.53		0.82		0.55		0.73
CEP-75%		0.41		0.27		0.53		0.81		0.86		0.43
CEP-90%		0.01		0.037		0.014		0.82		0.37		0.97
CEP-95%		<0.001		<0.001		0.003		0.63		0.97		0.95
CEP-99%		<0.001		<0.001		0.007		0.015		0.13		0.12



(a) Observed and fitted number of cases from negative binomial model B

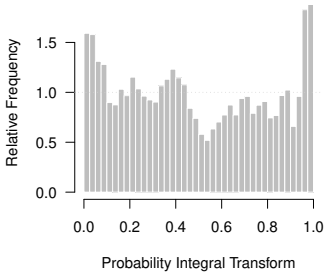
(b) Observed and fitted number of cases from negative binomial model C

Fig. 3: Observed and fitted number of meningococcal disease cases from negative binomial model B and C.

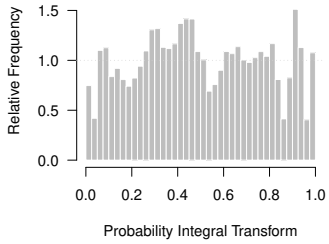
To assess the predictive performance of the models, the time series is divided into two parts: a learning set (2001–2002) and a validation set (2003–2006). 208 one-step-ahead predictions are calculated for the validation set in each model. We note that ML estimates of the model parameters have been re-estimated for each one-step-ahead prediction.

Table 3 shows mean scores and p-values from the proposed calibration tests for the different models. The mean scores generally prefer the negative binomial model with smaller values than for the Poisson model. Small p-values are reported for the Poisson model for most of the calibration tests, with the exception of some CEP tests, see Table 3. This indicates that all Poisson models produce miscalibrated forecasts. Only for large quantiles ($p = 90\%$, 95% and 99%), the CEP tests give some evidence against the null hypothesis that the Poisson models are well calibrated. In contrast, most of the negative binomial model are not rejected using a significance level of 0.05. The only exception is the CEP test based on the 99%-quantile applied to model A. For models B and C, the more realistic model allowing for autocorrelation, all CEP test produce insignificant p-values.

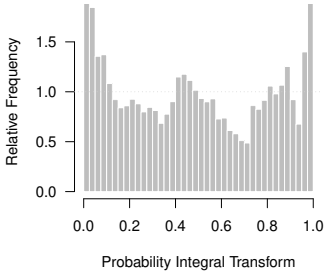
The non-randomised PIT (Czado et al, 2009) histogram shown in Figure 4 indicates that the predictions from all Poisson models are underdispersed with a typical U-shape. For count data, the underlying PIT values are no longer uniform under the null hypothesis of an ideal forecast. Therefore, a Kolmogorov test for PIT values can not be applied to check the calibration. A randomised PIT has been suggested based on a standard uniform random variable (Smith 1985). However, this approach is largely dependent on the random variable involved, which makes the test result not reliable. For example, for the Poisson



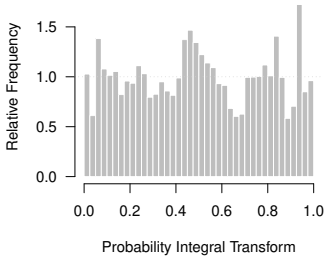
(a) Poisson Model A



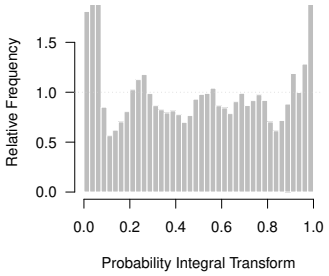
(b) Negative Binomial Model A



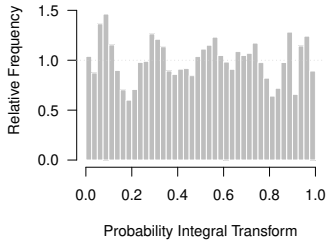
(c) Poisson Model B



(d) Negative Binomial Model B



(e) Poisson Model C



(f) Negative Binomial Model C

Fig. 4: Histogram for Probability Integral Transform (PIT) for model A, B and C based on Poisson (left) and negative binomial (right) distribution.

model C, we conducted 1000 tests based on the randomised PIT values with different seeds to generate the random variables. The p-values varied from 0.033 to 0.292. In conclusion, the calibration tests give no evidence that the negative binomial model B and C are miscalibrated. The negative binomial model A performs similarly, only the CEP test based on the 99%-quantile indicates some lack of calibration in the upper tail of the distribution. This may correspond to too many observations in this extreme quantile of the forecast distribution, compare the corresponding PIT histogram in Figure 4.
An interesting feature of the results shown in Table 3 is that the Poisson model C and the negative binomial model A have the same mean RPS score of 2.26. However, the corresponding p-value of the unconditional calibration RPS test identifies the Poisson model as strongly miscalibrated ($p < 0.001$), whereas there is no evidence of miscalibration of the negative binomial model ($p = 0.83$). Proper scoring rules incorporate both sharpness and calibration, so these results suggest that Poisson model C produces better forecasts in terms of sharpness. Indeed, the mean variance of the forecasts, a commonly used measure of sharpness, is 10.4 for the Poisson model C and 17.5 for the negative binomial model A. This reflects the more elaborate model structure of model C compared to model A. The negative binomial model A accommodates poor sharpness by overdispersion, making the forecasts well calibrated.

5 Discussion

In this paper, we have proposed several significance tests to assess calibration of predictive models for count data. In particular, we extend three different types of tests from continuous probabilistic predictions: the CEP test, and the unconditional and the regression tests based on proper scoring rules. Simulation results show all three types of tests are under good control of type I error, and are powerful tools for detecting the miscalibrated forecasts. Moreover, tests based on proper scoring rules are also powerful even if the number of observations is low. The application to data on meningococcal disease incidence illustrates that the calibration tests are a useful tool to detect miscalibration of forecasts.

These tests can be easily implemented, and can be applied in either a Bayesian or classical frequentist setting for forecast evaluation. They are used diagnostically to identify the deficiencies in calibration and furthermore facilitate model comparison and selection. Proper scoring rules are proved to be effective in evaluating both calibration and sharpness simultaneously (Gneiting et al, 2007). This partly explains why tests based on proper scoring rules also perform better than the CEP tests in this study. Among tests based on proper scoring rules, the regression approach for continuous predictions shows superior results than the unconditional tests to detect miscalibration (Held et al, 2010). However, for count data, it turns out to be the other way around, and unconditional tests seem to work better, except for overdispersed predic-

tions (see Figure 2b). Generally, CEP tests based on extremely large or small quantiles (e.g. 1% or 99%) are not recommended due to convergence problems of logistic regression. Among the tests based on proper scoring rules, the unconditional and the regression test based on DSS are recommended since they are simple to compute and performs almost best in simulation study.

The underlying assumption for all the tests in this paper is that the forecasts are independent. For unconditional tests based on proper scoring rules, the approximate distribution of statistics is based on the central limit theorem which requires the independence of random variables. Regression tests based on either proper scoring rules or CEPs assume independence of the residuals. A possible area of further research is to extend the tests to dependent forecasts. One possible direction could be to estimate the correlation structure of forecast errors and to incorporate it into the hypothesis tests (Diebold and Mariano, 1995).

We now discuss a modification of the unconditional test, which may perform better for low counts. For illustration, consider a series of Poisson forecasts with small rates μ_i for $i = 1, \ldots, n$. If μ_i is small, the variance $\mathsf{Var}_0(\text{DSS}) = 1 + 1/\mu_i$ of the DSS will become quite large, and the denominator of Z_s in (5) will be dominated by the forecasts with small rates μ_i. The alternative test statistic

$$Z_s^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{s_i - \mathrm{E}_0(s_i)}{\mathrm{Var}_0(s_i)^{1/2}},$$

may then perform better. Based on the central limit theorem of Liapounov, Z_s^* also approximately follows a standard normal distribution for large n.

Note that Z_s and Z_s^* are equivalent if $\mathsf{Var}_0(s_i)$ is the same for all predictions. This is the case for LS and DSS based on normal predictions, where $\mathsf{Var}_0(\text{LS}) = \mathsf{Var}_0(\text{DSS}) = 1/2$ (Held et al, 2010). However, for count data, the two test statistics are in general not equivalent and Z_s^* may be a more robust statistic than Z_s, especially in the presence of predictions with small rates.

We compared the power of both tests in a simulation study of 1,000 datasets with small rates. Here, the rate μ_i of the predictive distribution is a realization of a gamma $\mathrm{G}(1, 0.5)$ random variable with mean 2 and variance 4, and we set $\psi_i = \mu_i$ for the corresponding negative binomial predictions $\mathrm{NBin}(\mu_i, \psi_i)$. From Table 4 we see that the two test statistics Z_s and Z_s^* using RPS and LS have similar power. However, the test based on Z_{DSS}^* is more powerful than the one using Z_{DSS}. In this case, we would recommend to use the unconditional tests based on Z_s^*.

Finally, it is also of interest to know how much of the methods discussed in the paper can be generalised to multivariate forecasts. It is well-known that assessing calibration of multivariate forecasts is particularly challenging. For continuous predictions, calibration test based on LS can be conducted since the first two moments of the LS can be derived. However, the multivariate RPS

Test Statistic	Poisson					Negative binomial				
	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 500$	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 500$
Z_s										
RPS	0.478	0.699	0.922	0.937	0.995	0.100	0.100	0.106	0.127	0.335
LS	0.438	0.629	0.870	0.914	0.988	0.111	0.120	0.124	0.162	0.450
DSS	0.315	0.427	0.642	0.776	0.890	0.071	0.076	0.073	0.072	0.170
Z_s^*										
RPS	0.460	0.664	0.896	0.920	0.988	0.122	0.135	0.152	0.202	0.570
LS	0.444	0.636	0.871	0.918	0.988	0.108	0.127	0.146	0.191	0.541
DSS	0.400	0.578	0.841	0.911	0.982	0.106	0.137	0.157	0.228	0.615

Table 4: Power of two types of unconditional tests based on Z_s and Z_s^* for low counts.

is difficult to explore (Gneiting et al, 2008). Calibration test of multivariate forecasts of count data will be even more challenging, especially due to the complexity to compute the expectation and variance under the null hypothesis.

Acknowledgements

We thank two referees for helpful comments and suggestions. Financial support by the Swiss National Science Foundation (SNF) is gratefully acknowledged.

References

Brier GW (1950) Verification of forecasts expressed in terms of probability. Monthly Weather Review 78:1–3

Christoffersen PF (1998) Evaluating interval forecasts. International Economic Review 39(4):pp. 841–862

Corradi V, Swanson NR (2006) Predictive density and conditional confidence interval accuracy tests. Journal of Econometrics 135(1):187–228

Cox DR (1958) Two further applications of a model for binary regression. Biometrika 45:562–565

Czado C, Gneiting T, Held L (2009) Predictive model assessment for count data. Biometrics 65:1254–1261

Dawid AP (1984) Statistical theory: the prequential appoach. Journal of the Royal Statistical Society, Series A 147:278–292

Dawid AP, Sebastiani P (1999) Coherent dispersion criteria for optimal experimental design. Annals of Statistics 27:65–81

DeGroot M, Schervish M (2012) Probability and Statistics, 4th edn. Addison-Wesley, Boston

Diebold FX, Mariano RS (1995) Comparing predictive accuracy. Journal of Business & Economic Statistics 13(3):pp. 253–263

Diebold FX, Gunther TA, Tay AS (1998) Evaluating density forecasts with applications to financial risk management. International Economic Review 39(4):pp. 863–883

Elsner JB, Jagger TH (2006) Prediction models for annual US hurricane counts. Journal of Climate 19(12):2935–2952

Epstein ES (1969) A scoring system for probability forecasts of ranked categories. Journal of Applied Meteorology 8:985–987

Farrington CP, Andrews NJ, Beale AD, Catchpole MA (1996) A statistical algorithm for the early detection of outbreaks of infectious disease. Journal of the Royal Statistical Society, Series A 159:547–563

Frühwirth-Schnatter S, Frühwirth R, Held L, Rue H (2009) Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. Statistics and Computing 19(4):479–492

Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society, Series B 69:243–268

Gneiting T, Stanberry LI, Grimit EP, Held L, Johnson NA (2008) Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. Test 17(2):211–235

Good IJ (1952) Rational decisions. Journal of the Royal Statistical Society, Series B 14:107–114

Harvey DI, Leybourne SJ, Newbold P (1998) Tests for forecast encompassing. Journal of Business & Economic Statistics 16(2):pp. 254–259

Heisterkamp SH, Dekkers AL, Heijne JC (2006) Automated detection of infectious disease outbreaks: hierarchical time series models. Statistics in Medicine 25(24):4179–4196

Held L, Paul M (2012) Modeling seasonality in space-time infectious disease surveillance data. Biometrical Journal 54(6):824–843

Held L, Höhle M, Hofmann M (2005) A statistical framework for the analysis of multivariate infectious disease surveillance counts. Statistical Modelling 5:187–199

Held L, Hofmann M, Höhle M, Schmid V (2006) A two-component model for counts of infectious diseases. Biostatistics 7(3):422–437

Held L, Rufibach K, Balabdaoui F (2010) A score regression approach to assess calibration of continuous probabilistic predictions. Biometrics 66(4):1295–1305

Hilbe JM (2011) Negative Binomial Regression, 2nd edn. Cambridge University Press, Cambridge

Katti S (1960) The moments of the absolute difference and the absolute deviation of distributions. The Annals of Mathematical Statistics 31:78–85

Knessl C (1998) Integral representations and asymptotic expansions for Shannon and Renyi entropies. Applied Mathematics Letters 11(2):69–74

Manitz J, Höhle M (2013) Bayesian outbreak detection algorithm for monitoring reported cases of campylobacteriosis in Germany. Biometrical Journal

Mason S, Galpin J, Goddard L, Graham N, Rajartnam B (2007) Conditional exceedance probabilities. Monthly Weather Review 135(2):363–372

McCabe B, Martin G (2005) Bayesian predictions of low count time series. International Journal of Forecasting 21(2):315 – 330

McCabe BP, Martin GM, Harris D (2011) Efficient probabilistic forecasts for counts. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73(2):253–272

Murphy AH, Winkler RL (1987) A general framework for forecast verification. Monthly Weather Review 115:1330–1338

Nelson K, Leroux B (2006) Statistical models for autocorrelated data. Statistics in Medicine 25:1413–1430

Noufaily A, Enki DG, Farrington P, Garthwaite P, Andrews N, Charlett A (2013) An improved algorithm for outbreak detection in multiple surveillance systems. Statistics in Medicine 32(7):1206–1222

Paul M, Held L, Toschke A (2008) Multivariate modelling of infectious disease surveillance data. Statistics in Medicine 27:6250–6267

Smith JQ (1985) Diagnostic checks of non-standard time series models. Journal of Forecasting 4:283–291

Spiegelhalter DJ (1986) Probabilistic prediction in patient management. Statistics in Medicine 5:421–433

Steyerberg E (2009) Clinical Prediction Models. Springer, New York

Winkelmann R (2008) Econometric Analysis of Count Data, 5th edn. Springer, New York

Winkler RL (1996) Scoring rules and the evaluation of probabilities. Test 5(1):1–60